# Image classification for Web genre identification

Alex Wu, Naval Research Laboratory, Code 5584, *Summer Intern*
Myriam Abramson, Naval Research Laboratory, Code 5584

***Abstract*— With the countless number of existing websites alongside the virtually unrestricted growth of the World Wide Web, the Web has no boundaries. As a result, there is an increasing need to automatically categorize and classify web sites into genres in order to improve the personalization of search results. This paper will offer conceptual suggestions on how online images can be used to predict the genre of the website that they are found on, as well as the process for detecting and identifying certain specific images for genre categorization.**

## [1] Introduction

Several approaches to genre classification have been proposed. In [2] the authors investigated the categorization of websites into genres based on mnemonics found in the URL (e.g. having "wordpress" in the URL would highly suggest a blogging site). However, there are several other possible ways in which genre classification can be conducted, including images found on the website. This paper will offer conceptual suggestions on how online images can be used to predict the genre of the website that they are found on, as well as the process for detecting and identifying certain specific images, which can then be used to categorize a user's general web browsing activity and habits.

There are already a sizable number of papers available on web genre classification, several in regards to improving search engine accuracy. In addition to uses for search engines, however, genre classification also helps categorize web browsing behavior into presets, which subsequently improves the ease of matching web histories with each other [1]. Unlike genres of other forms, web genres are analyzed more based on their style, layout, or formatting rather than purely on content [3]. An example of traditional use of genres can be seen in literature, where genres are based purely on the content/plot of the book. While it is certainly still possible to use this content based method online, layout based classification is also very viable [4]. Unlike books, webpages of different genres generally appear differently from one other. For example, a site that sells products (Ebay or Amazon) looks vastly different in terms of their layout in comparison to a site of a different genre (say forums/discussions). Although there can be a virtually infinite number of genres for online classification, having too many genres would defeat the overall purpose of having general categorizations, whereas too few would not accurately represent the majority of websites. However, there are some genres that can generally be applied for web use. These include blogs, information sites, corporate homepages, personal homepages, discussion/forums, news sites, and online shops. Of course overlap is bound to occur (an online shop's homepage is also its corporate homepage), but those are a few to start out with. The text of a webpage has been used to provide contextual features for image classification [5]. In this paper, we

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **2012** | | **00-00-2012 to 00-00-2012** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Image classification for Web genre identification** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Research Laboratory,Code 5584,4555 Overlook Ave., SW,Washington,DC,20375** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

**With the countless number of existing websites alongside the virtually unrestricted growth of the World Wide Web, the Web has no boundaries. As a result, there is an increasing need to automatically categorize and classify web sites into genres in order to improve the personalization of search results. This paper will offer conceptual suggestions on how online images can be used to predict the genre of the website that they are found on, as well as the process for detecting and identifying certain specific images for genre categorization.**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **Same as Report (SAR)** | **5** | |

propose to use image types as features for genre classification in webpages. In order to use image content for genre classification, a database of websites of a certain genre must be analyzed and similarities in image content should be noted. For corporate homepages, there will always be a logo present (usually in the upper part of the site). E-shop pages might contain the logo of a credit card. Download pages might contain an arrow icon to start the download. In personal homepages, a personal photo of the writer is often present. Extending on personal homepages, a personal travel blog will most likely contain miscellaneous photographs as well as scenic landscapes of the writer's travels. Although there is a lot of research necessary to be conducted to create and confirm patterns of image content in websites, these are some preliminary examples for experimentation purposes. In regards to classifying images (as faces or logos etc.), one would think there may be possibilities with web analysis of the images (such as ALT tags). To test the viability of ALT tags, we tested the percentage of images that had ALT tags on 5000 websites and found that only 40% of the images contained in those URLs were ALT tagged. This percentage is an upper bound on the accuracy that could be obtained from image classification according to ALT tags. Additionally, processing ALT tags could also prove difficult; if we assumed that a sizable portion profile pictures were labeled with the subjects name and corporate logos were labeled with the corporation name, there would be no simple method for connecting these names with a face/personal photo and logo respectively.

## [2] METHODOLOGY

For this project, we focused primarily on facial recognition and landscape detection using the computer vision toolkit OpenCV[1]. For facial recognition, we researched the possibilities of using the very popular method of Haar classifiers. Haar-like features are obtained by comparing adjacent rectangular areas in the image and finding the difference of pixel intensities between the regions [6]. Figure 2 showcases several basic Haar-like features. Examples of how these features are used include the difference in pixel intensities between the eyes and upper cheek in a face (the eyes have a higher intensity than the cheeks) and the difference between the nose bridge and cheeks. A visual example created by Paul Viola and Michael Jones is shown in Figure 1 and 2 below.

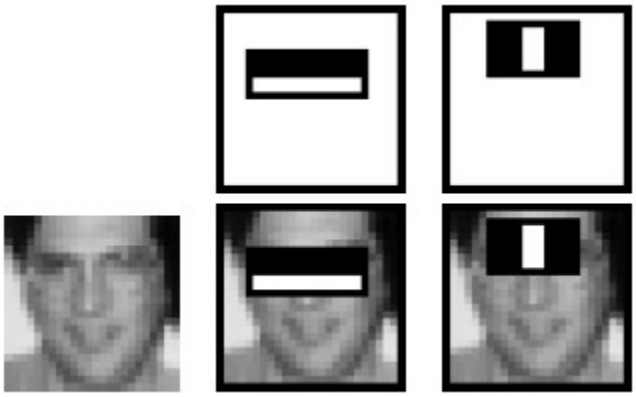
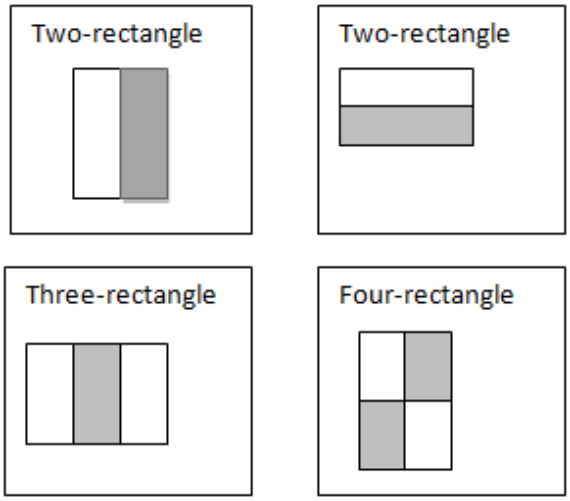


*Figure  Haar-like features for face recognition [6]*

1 http://opencv.org




*Figure*

The problem with Haar cascades is that they rely on very rigid image models; if the angle or viewpoint of the test image differs from the images used to train the classifier, the subject will not be detected. As a result, we decided to stick with general front facing faces (which are mostly used for profile pictures on websites). For the experiment, we simply used the default Haar face cascades (created by Raine Lienhart) that came pre-packaged with OpenCV.

Landscape detection, on the other hand, is vastly different from facial detection. Whereas all faces (from the front view) have the same general structure and difference in shading, landscapes are extremely varied in their image structure. Consequently, Haar classifiers were far out of the question for identification. Instead, we decided to focus more on the green and blue colors of landscapes (landscapes with other colors such as sunsets were excluded from the experiment for simplicity sake). In order to use colors for landscape detection, we created a small directory of 36 positive landscapes and negative images and compared their image histograms with 50 test images (25 positive 25 negative) using the correlation method where H is the matrix histogram of the hue and saturation of the images and N is the number of bins:

The image is then determined to be or not be a landscape based on a nearest neighbor classification where k=15. The graph below charts the results of testing an image dataset of 25 positives and negatives (assembled specifically for determining the k coefficient), and 15 was found to have the best balance between high positive hit rates and low false positives (Fig. 3).

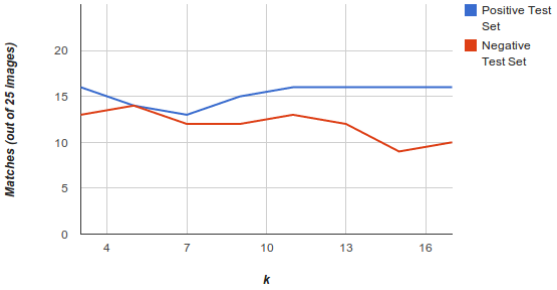
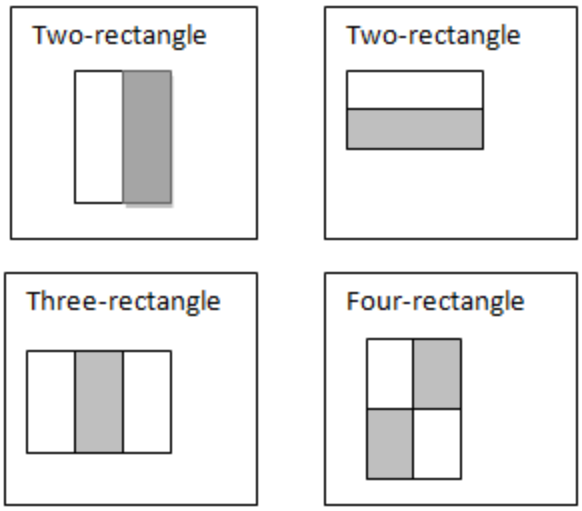



*Figure K nearest-neighbor outcomes*

We additionally added another factor to determine positive landscapes; the average correlation between the test image and the images in the positive training set had to be at least 0.1. Similarly to the k-coefficient training, we conducted a test using the positive dataset of images used for the k training to determine the minimum average correlation (Fig 4).



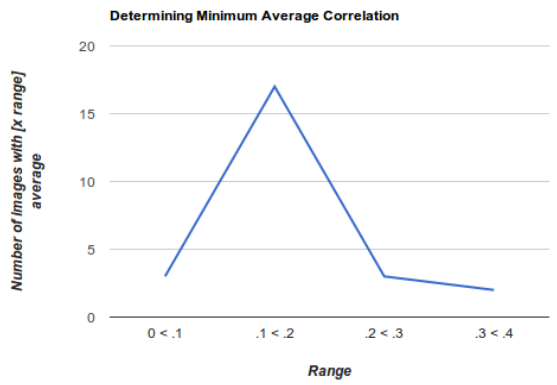



*Figure Minimum average correlation of images in test set and training set*

## [3] EXPERIMENTATION

To test the OpenCV face Haar cascades (frontalface_default, frontalface_alt, and frontalface_alt2), we downloaded an old set of 100 profile images shot from the Olivetti Research Laboratory in Cambridge, UK to ensure a high percentage of faces were detected along with a set of 100 random negative images (these range from landscapes to interior pictures and random objects) to test the false positive rate of faces occuring. Although a larger test set should certainly be used, we simply wanted a overall preliminary test to do a basic comparison between the cascades. The results are as follows:

| Matches (out of 200) | Cascade 1 | Cascade 2 | Cascade 3 |
|---|---|---|---|
| #True Positives | 83 | 86 | 88 |
| # False Positives | 42 | 20 | 24 |
| %Correct | 70.5 | 83 | 82 |

As the results show, the first cascade is certainly not the best choice based on the test, and neither the second nor third classifier have a significant lead over another, it is more of a matter of whether or not you prefer positive accuracy on faces or a low hit on false positives. The number of positive faces missed was most likely caused by either the face in question being at a slightly off angle position, or possibly the subject wearing glasses. In order to make this face detector slightly more applicable to web genre classification (many images outside of profile pictures may include a picture of a face), we set several experimental measures, the most important one being that the detected rectangular face area must equal at least 1/5 of the area of the entire image, eliminating images where a detected face is not the primary focus of the picture.

For the landscape detection program, we also conducted a very small test to get basic results. Using a new image test set consisting of 50 images (with 25 positive images and 25 negative images) with k set to 15 and minimum average correlation to .1

(based on conducted test runs), results were as follows:

| Matches (out of 50) | Test Set |
|---|---|
| #True Positives | 20 |
| #False Positives | 5 |
| %Correct | 80 |

## [4] Conclusion

Although more extensive tests must obviously be conducted, the preliminary results suggest that there is potential for these image classifiers to be used to detect their respective targets (faces and landscapes), which can then be used for the purpose of genre classification. Additionally, there are many other possibilities for genre classification with images, such as with logos (which can be used to detect non-personal homepages as mentioned before). However, the generally unpredictable nature of a logo in both colors and shape make it difficult if not near impossible to classify based on the methods we applied for faces and landscapes.

## References


[1] Abramson M., Toward the Attribution of Web Behavior, IEEE Symposium on Computational Intelligence for Security and Defence Applications, 2012.

[2] Abramson M., Aha D.,What's in a URL? Genre Classifications from URLs, ITWP Workshop at AAAI, 2012.

[3] Boese E., Stereotyping the Web: Genre Classification of Web Documents, Master's Thesis, Colorado State University, 2005.

[4] Levering R., Using Visual Features for Fine-Grained Genre Classification of Web Pages, Hawaii International Conference on System Sciences, 2008.

[5] Kavla P., Enembreck F., Koerich A., WEB Image Classification Based on the Fusion of Image and Text Classifiers, Int'l Conference on Document Analysis and Recognition, 2007.

[6] Viola P., Jones M., Rapid Object Detection using a Boosted Cascade of Simple Features, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.